Final Report for Period: 09/2007 - 08/2008 **Submitted on:** 11/30/2008

Principal Investigator: Lee, Eva K. **Award ID:** 0300435

Organization: GA Tech Res Corp - GIT

Submitted By:

Lee, Eva - Principal Investigator

Title:

Investigations in Combinatorial Optimization and its Applications to DNA Sequencing Problems

Project Participants

Senior Personnel

Name: Lee, Eva

Worked for more than 160 Hours: Yes

Contribution to Project:

Name: Johnson, Ellis

Worked for more than 160 Hours: No

Contribution to Project:

Post-doc

Graduate Student

Name: Gupta, Kapil

Worked for more than 160 Hours: Yes

Contribution to Project:

Kapil Gupta worked primarily on the DNA sequencing problem as part of his Ph.D. thesis research. Under the supervision of Lee, he was learning the biological implications of DNA changes in methylation and human cancer, and conducted theoretical and computational investigation. A PhD thesis was completed in 2008.

Name: Zhang, Yang

Worked for more than 160 Hours: No

Contribution to Project:

This individual was learning optimization approaches for biological problem that was proposed in this project.

Name: Shi, Jin

Worked for more than 160 Hours: No

Contribution to Project:

Jin Shi learnt of the pattern recognition algorithm developed in this research and performed empricial experiments on cancer prediction.

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Dr. Joe Costello, University of California at San Francisco, Comprehensive Cancer Center, has supplied us with some brain tumor DNA sequences. We applied our pattern search and combinatorial approach to analyse the characteristics of the sequences, and their possible differences from DNA sequences from normal individuals.
Dr. Paula Vertino, Emory University, has supplied us with DNA sequences related to breast cancer and lung cancer for analysis.

Activities and Findings

Research and Education Activities:

Research:

DNA comparison and evolutionary biology have been referred to as the cornerstones of biology. Evolutionary explanations of biological phenomena pervade all fields of biology and bring them together under one theoretical umbrella. DNA sequencing problems in particular have been extensively studied, and there is a tremendous body of literature describing different models and their analysis under different metrics of uncertainty, as well as fast heuristic approaches for solving model instances. Due to the NP-hard nature of these problems and to their fundamental importance in biology, further research in this area is imperative. In this study, we focus on a combinatorial approach that enables one to view the evolutionary distance problem both as an integer program and as a node-packing problem. This approach opens up avenues for developing sophisticated exact and heuristic algorithms based on the well-developed theories and methods of integer programming and node-packing.

A cancer biology problem that provides motivation for our research involves DNA methylation patterns and their link to human cancer. Methylation is a process in which methyl groups (CH_3) are added to the bases (A,T,G,C) that constitute DNA. Aberrant methylation of normally unmethylated stretches of the genome (CpG islands) occurs frequently in human cancers, and is associated with the inactivation of key tumor suppressor and other genes. At present, it is not known how or why particular genes succumb to methylation-mediated silencing. We have applied a variant of our combinatorial approach to identify patterns in stretches of the human genome that can help to predict methylation status. This aspect of our research is of interest to the large number of biologists studying this important gene inactivation mechanism, to medical oncologists who treat cancer patients, and ultimately to cancer patients themselves.

Education:
Supervise 3 Ph.D. students, each was involved in various ways in this project.

Presentation:
EK Lee, Combinatorial and integer programming techniques for DNA sequencing problems, INFORMS Hawaii, June 2001.

EK Lee, Clique Inequalities & their Use in Solving DNA Sequencing Problems, INFORMS Miami, Nov 2001

EK Lee, Bioinformatics and Gene expression analysis, 11th Annual Suddath Symposium and Annual Georgia Cancer Coalition Spring Symposium 2003, Mar 2003.

EK Lee, Large-Scale Genomic Pattern Recognition and Classification of Hypermethylated CpG Loci in Human Cancer, 11th Annual Suddath Symposium and Annual Georgia Cancer Coalition Spring Symposium 2003, Mar 2003.

EK Lee, Cancer Bioinformatics: Cutting-edge Cancer Technologies at Georgia Tech and Emory University, Presentation to Georgia Cancer Coalition, May 13 2002.

P Vertino, Epigenetic silencing in Human Cancer: Lessons from TMS1,
Department of Biochemistry and Molecular Biology, University of Florida, Gainseville FL, August 17, 2003.

EK Lee, Pattern Recognition and Discriminant Analysis in Genomic Analysis of Human Cancer, INFORMS Atlanta, October 2003.

EK Lee, Bioinformatics, Modeling, and Biocomputing, Emory-GA Tech Planning Symposium on Quantitative Medicine, Dec 2003.

EK Lee, Discriminant Analysis and Predictive Models in Medicine,
University of Georgia Computational Biology Seminar, Jul 2004.

EK Lee, Large-Scale Genomic Pattern Recognition and Classification of CpG Loci in Human Cancer, Center for Bioinformatics and Computational Biology (BiComB) Seminar, Sep 2004.

EK Lee, Large-Scale Genomic Pattern Recognition and Classification of
CpG Loci in Human Cancer, INFORMS Colorado, October 2004.

Discriminant Analysis and Predictive Models in Medicine,
SouthEast Collaborative Alliance Biocomputing Center (SECABC) Winter Workshop on Biocomputing, Atlanta Jan 2005.

Large-Scale Genomic Pattern Recognition and Classification of CpG Loci in Human Cancer, INFORMS San Francisco Nov 2005.

Large-Scale Biocomputing for Cancer Diagnosis, INFORMS San Francisco Nov 2005.

Gupta, Novel Evolutionary Models and Applications to the Sequence Alignment Problem. The Fifth Georgia Tech - Oak Ridge International Conference on Bioinformatics -- Computational Genomics and Evolutionary Biology,
Nov 2005.

Large-Scale Genomic Pattern Recognition and Classification of Aberrantly Hypermethylated CpG Island Loci in Human Cancer. The Fifth Georgia Tech - Oak Ridge International Conference on Bioinformatics --Computational Genomics and Evolutionary Biology, Nov 2005.

Pattern Recognition and Classification in Medical Diagnosis and Disease Prediction, INFORMS Pittsburgh Nov 2006.

Distinguished presentation: Discriminant Analysis and Predictive Models in Medicine and Biology,Emory Computational and Life Sciences (CLS) Strategic Initiative, Mar 2007.

Gupta, Integer Programming Models for DNA Sequencing Problems, INFORMS Seattle 2007.

Gupta, A Large-scale Computational Model for DNA Sequencing Problems, INFORMS Washington 2008.

Findings:

We proved that the minimum weight common mutated sequence (MWCMS) is in NP, and when the number of input sequences is a constant, MWCMS is polynomial-time solvable. We also established that some well-known combinatorial problems can be viewed as special cases of the MWCMS problem. These include the minimum weight supersequence problem, the minimum weight superstring problem, and the longest common subsequence problem. Furthermore, the conflict graph constructed from the complete paths of the multi-layer graph is perfect, and thus the corresponding node-packing problem is polynomial-time solvable. However, for practical situations, the number of complete paths is huge. We applied a column-generation approach to solve the entire node-packing problem. This is exciting as it opens up a new way to tackle the evolutionary distance problem, and offers advances in the computational strategies for solving large-scale node-packing problems. (Node packing is a classical problem studied in combinatorial optimization, and is encountered in the study of many important applications.) Computational strategies that involve a simultaneous column and row generation approach was explored. To the best of our knowledge, column generation has not previously been applied to solve node-packing problems.

Heuristic approaches identified hundreds of short patterns that could potentially be useful for predicting methylation status of CpG islands in human cancer. Applying the patterns to the CpG islands provided by medical collaborators Vertino and Costello, a set of seven sequence attributes were selected and used to develop prediction rules that can classify a set of CpG islands with unknown status with an overall accuracy of 82%. This is extremely exciting as the results suggest that there may be a sequence signature associated with aberrant DNA methylation. An important aspect of the medical collaboration is that the cancer biologists can now validate the prediction by testing in their laboratory the methylation status of the ``unknown'' CpG islands, providing a strong feedback mechanism that will improve the accuracy and thus the importance of the computational work. Furthermore, two sequence patterns detected by our algorithm are known biological entities (ancient retroviral elements) allowing cancer biologists to develop hypotheses regarding the potential role of these sequences in aberrant methylation, an event occurs frequently in human cancer.

Training and Development:

The project provided Ph.D. and undergraduate students hands-on research experience in interdisciplinary areas of engineering and bioinformatics. Each student works closely with the PI as well as well-known biological researchers. Three PhD students involved in this project took classes in their main engineering disciplines as well as those in biology and bioinformatics disciplines. They were also working towards a minor in bioinformatics during their graduate studies. The students also had the opportunities to attend conferences and present results of their findings. A PhD thesis was completed in 2008.

Outreach Activities:

The PI is involved in high-school recruiting and has presented her work in interdisciplinary areas to some of the high-schoolers.

The PI presented engineering techniques and their applications to biological applications at the 'Georgia Tech: Innovating Here and Now', a public alumni event (approx. 300 attendees from very diverse fields) held in Washington DC, November 2004.

The PI was selected as the 2006 Mathematician Ambassador by the American Mathematical Society in June 2006 to meet individually with congressional leaders on Capital Hill to present and promote mathematical advances and applications to biology and medicine, and to lobby federal funding increase to NSF.

The PI gave the keynote presentation on Research Challenges in Medicine and HealthCare in the 31st Conference on Math ematics of Operations Research, Lunteren, The Netherlands Jan 2006.

Journal Publications

FA Feltus, EK Lee, JF Costello, C Plass, PM Vertino, "Predicting Aberrant CpG Island Methylation", Proceedings of the National Academy of Sciences, p. 12253, vol. 100(21), (2003). Published,

EK Lee, T Easton, EL Johnson, " Combinatorial Approach to the Minimum Weight Common Mutated Sequence Problem and its Application to DNA Sequencing Problems", Georgia Tech Technical Paper, p. , vol. , (2004). Published,

EK Lee, K Gupta, T Easton, "Novel Evolutionary Models and its Application to Sequence Alignment Problem", Annals of Operations Research, p. 16, vol. 14, (2006). Published,

FA Feltus, JF Costello, C Plass, EK Lee, PM Vertino, "Identifying sequence patterns associated with aberrant CpG island methylation", Proc. Amer. Assoc. Cancer Res., p. LB:3, vol. , (2002). Published,

Feltus, FA; Lee, EK; Costello, JF; Plass, C; Vertino, PM, "DNA motifs associated with aberrant CpG island methylation", GENOMICS, p. 572, vol. 87, (2006). Published, 10.1016/j.ygeno.2005.12.01

McCabe MT, Lee EK, Vertino PM, "A Multi-Factorial Signature of DNA Sequence and Polycomb Binding Predicts Aberrant CpG Island Methylation", Cancer Research, p. , vol. , (2009). Accepted,

Lee, EK, "Optimization-Based Predictive Models in Medicine and Biology", Optimization in Medicine, Springer Series in Optimization and Its Application, p. 127, vol. 12, (2007). Published,

Lee, EK, " Large-scale optimization-based classification models in medicine and biology", Annals of Biomedical Engineering, Systems Biology and Bioinformatics, p. 1095, vol. 35(6), (2007). Published,

Books or Other One-time Publications

EK Lee, K Gupta, "Algorithms for Genomics Analysis", (2007). Book, Accepted

Editor(s): Chris A. Floudas, and Panos M. Pardalos
Collection: Encyclopedia of Optimization
Bibliography: Springer, The Netherlands

EK Lee, K Gupta, "Algorithms for Genomics Analysis", (2007). Book, Accepted
Editor(s): Optimization in Medicine
Collection: E Romeijn and PM Pardalos
Bibliography: Kluwer Academic Publishers

EK Lee, K Gupta, "Algorithms for Genomics Analysis", (2007). Book, Published
Editor(s): G. Lim and EK Lee
Collection: Optimization in Medicine and Biology
Bibliography: Taylor & Francis Group, LLC, CRC Press

Gupta, K, "Combinatorial Optimization and Application to DNA Sequence Analysis", (2008). Thesis, Published
Bibliography: August

Web/Internet Site

URL(s):

http://www.isye.gatech.edu/~evakylee/medicalor

Description:

This website features biomedical research projects conducted by the PI, including this NSF-sponsored project. It also includes inter-disciplinary graduate programs of Operations Research in Medicine, and Bioinformatics, as well as undergraduate premed programs. Funding from NSF is acknowledged.

Other Specific Products

Contributions

Contributions within Discipline:

The mathematical model is useful in theoretical studies of difficult combinatorial problems; it also offers a novel way to tackle the DNA comparison & evolutionary distance problems -- cornerstones of biology.

Computational development for MWCMS advances knowledge in large-scale integer programming and node-packing problems, which arise frequently in industrial applications. Furthermore, the study advances the field of computational biology.

Our research effort has established a novel graph-theoretical approach for representing a wide variety of genomic sequence analysis problems within a single model. The model allows incorporation of the operations 'insertion', 'deletion', and 'substitution', and various parameters such as relative distances and weights. Conceptually, we refer the problem as the minimum weight common mutated sequence (MWCMS) problem. The MWCMS model has many applications including multiple sequence alignment problem, the phylogenetic analysis, the DNA sequencing problem, and sequence comparison problem, which encompass a core set of very difficult problems in computational biology. Thus our work lays out a mathematical modeling framework that allows one to investigate theoretical and computational issues, and to forge new advances for these distinct, but related problems.

Contributions to Other Disciplines:

The research has a significant impact on early cancer detection, provides novel molecular targets for chemotherapeutic intervention treatment strategies, and identifies genomic methylation markers for cancer prediction, treatment and prognosis.

Specifically, the implications of these findings are several-fold. First, the identification of sequence patterns/attributes that distinguish methylation-prone CpG islands will lead to a better understanding of the basic mechanisms underlying aberrant CpG island methylation. Because genes that are silenced by methylation are otherwise structurally sound, the potential for reactivating these genes by blocking or reversing the methylation process represents an exciting new molecular target for chemotherapeutic intervention. A better understanding of the

factors that contribute to aberrant methylation, including the identification of sequence elements that may act to target aberrant methylation, will be an important step in achieving this long-term goal. Secondly, the classification of the more than 29,000 known (but as yet unclassified) CpG islands will provide an important resource for the identification of novel gene targets for further study as potential molecular markers that could impact on both cancer prevention and treatment. Extensive information on CpG island methylation events (and thus potential training sets of methylated CpG islands) already exist for many human tumor types, including breast, brain, lung, leukemias, hepatocellular carcinomas, and PNET . Thus, the computational method developed will be directly applicable to CpG island methylation data derived from human tumors. Thus, the research has the potential to lead to improved diagnosis, prognosis and treatment planning for cancer patients.

Contributions to Human Resource Development:

The project offers opportunities for Ph.D. and undergraduate students to work in interdisciplinary areas of engineering and bioinformatics. Each student works closely with the PI as well as well-known biological researchers. Three PhD students involved in this project all took classes in their main engineering disciplines as well as in biology and bioinformatics disciplines. They were also working towards a minor in bioinformatics during their graduate studies. A PhD thesis was completed in 2008, and another will be completed in 2009.

The PI is involved in high-school recruiting and has presented her work in interdisciplinary areas to some of the high-schoolers.

The PI presented engineering techniques and their applications to evolutionary biology at the 'Georgia Tech: Innovating Here and Now', a public alumni event (approx. 300 attendees from very diverse fields) held in Washington DC, November 2004.

Since 2005, the PI has been co-chairing the bi-annual 'International Conference on Bioinformatics -- In silico Biology' with top bioinformatician, Dr. Borodovsky. The PI was responsible for inviting over 20 world-class bioinformaticians for plenary presentation. The conference has attracted over 150 researchers from around the world to attend. The conference helps foster multi-disciplinary research collaboration among OR and life science researchers.

Contributions to Resources for Research and Education:

Three book chapters on relatd topics, Algorithms for Genomic Analysis were published.

Contributions Beyond Science and Engineering:

The project may contribute to the healthcare benefits of the society as a whole. Specifically, the research has a significant impact on early cancer detection, provides novel molecular targets for chemotherapeutic intervention treatment strategies, and identifies genomic methylation markers for cancer prediction, treatment and prognosis.

Our research effort has established a novel graph-theoretical approach for representing a wide variety of genomic sequence analysis problems within a single model. The model allows incorporation of the operations 'insertion', 'deletion', and 'substitution', and various parameters such as relative distances and weights. Conceptually, we refer the problem as the minimum weight common mutated sequence (MWCMS) problem. The MWCMS model has many applications including multiple sequence alignment problem, the phylogenetic analysis, the DNA sequencing problem, and sequence comparison problem, which encompass a core set of very difficult problems in computational biology. Thus our work lays out a mathematical modeling framework that allows one to investigate theoretical and computational issues, and to forge new advances for these distinct, but related and important biological problems.

Categories for which nothing is reported:

Organizational Partners

Any Product